International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 07, NCIRTAI - 2026

Multivariate Spatial Trees for Bayesian Regression in Big Data

Mrs. EL. Thanga Uma.,
Assistant Professor, Department of Information Technology,, Sri Bharathi Engineering College for Women, Pudukkottai.
uma085@gmail.com

Abstract - Standard geostatistical models based on Gaussian processes are known to not scale to huge data volumes, making high resolution geospatial data difficult to handle. approaches for large-scale data that enable the depiction of complex relationships between several outcomes recorded at high resolutions by various sensors have received much less attention than approaches that can be computed more efficiently. By assuming conditional independence on latent random effects after a treed directed acyclic graph, our spatial multivariate tree-based Bayesian multivariate regression models (SPaMTREES) are scalable. The building of the tree and the associated efficient sampling methods in imbalanced multivariate contexts are guided by information-theoretic reasoning and computing efficiency considerations. We demonstrate SPaMTREES utilising a sizable climate data collection that integrates satellite and land-based station data, in addition to simulated data instances.

Keywords--- Directed acyclic graph, Gaussian process, Geostatistics, Multivariate regres- sion, Markov chain Monte Carlo, Multiscale/multiresolution

I.INTRODUCTION

Gathering vast amounts of georeferenced data is becoming more and more prevalent in the scientific and social sciences. Through interpretable models that measure uncertainty while accounting for spatial and temporal dimensions, researchers hope to use these data to comprehend phenomena and make predictions. Gaussian processes (GP) are versatile tools that may be used to assess uncertainty and characterise temporal and geographical variability. Much effort has gone into creating GP-based techniques that get over their infamously poor scalability to massive data.

There is now a wealth of research on scaling GPs to big scales. We discuss low-rank methods (Quiñonero-Candela and Rasmussen, 2015; Snelson and Ghahramani, 2017; Banerjee et al., 2018; Cressie and Johannesson, 2018); their extensions (Low et al., 2015; Ambikasaran et al., 2018; Huang and Sun, 2023; Geoga et al., 2024); and methods that take advantage of special structure or simplify the representation of multidimensional inputs.In geostatistical scenarios, which concentrate on small-dimensional inputs, such as the spatial coordinates plus time, these methods could not be available or perform poorly. In these situations, Toeplitz-like structures are usually absent, low-rank approaches oversmooth the spatial surface (Banerjee et al., 2022), and so-called separable covariance functions produced by tensor products poorly characterise spatial and temporal dependence.

Multivariate (or multi-output) regression settings present additional challenges. Multivariate geostatistical data are frequently detected at non-overlapping geographical regions, or misaligned (Gelfand et al., 2010). A number of variables are measured at non-overlapping places in Figure 1, with one measurement grid being significantly sparser than the others. Predicting outcomes at new locations in these contexts can be accomplished by substituting distinct single-output models for a multi-output regression. Although single-output models may occasionally perform on par with or even better than multi-output models, they are unable to identify and quantify cross-dependencies between outputs; determining if such dependences exist may have a greater scientific impact than making predictions.

II. RELATED WORKS

In order to address these problems, we present a Bayesian regression model in this article that encodes spatial dependence as a latent spatial multivariate tree (SpamTree); all non-reference locations are assigned to leaf nodes of the same DAG using a map, while conditional independence relations at the reference locations are controlled by the branches in a treed DAG. This assignment map maintains the desirable recursive properties of treed DAGs while controlling the nature and size of the conditioning sets at all locations. It is used to improve estimation and predictions and overcome common problems in standard nearest-neighbor and recursive

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 07, NCIRTAI - 2026

partition methods when severe restrictions on the reference set of locations become necessary due to data size.

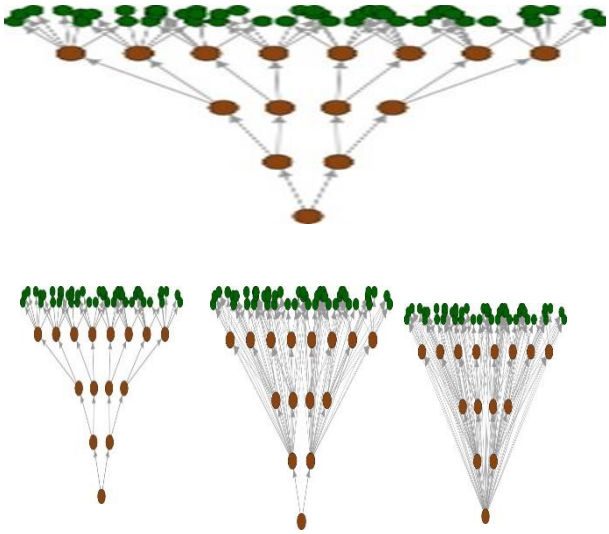

Figure 2: Three SpamTrees on $M = 4$ levels with depths $\delta = 1$ (left), $\delta = 3$ (center), and $\delta = 4$ (right). Nodes are represented by circles, with branches colored in brown and leaves in green.

By creating a technique that focuses on effective computations of Bayesian multivariate spatial regression models, the current study contributes to the expanding body of literature on spatial processes described on DAGs. While MRAs are defined as a basis function expansion, SpamTrees can be represented as a treed graph of a SpamTree with full "depth" as specified later (the DAG on the right of Figure 2), in univariate settings, and "response" models (Katzfuss, 2017).

III.SPATIAL MULTIVARIATE TREES

Consider a spatial or spatiotemporal domain D. With the temporal dimension, we have D ⊂ □d ×[0, ∞), otherwise D ⊂ □d. A q-variate spatial process is defined as an uncountable set of random variables {v(l) : l ∈ D}, where v(l) is a q × 1 random vector with elements wi(l) for i = 1, 2, . . . , q, paired with a probability law P defining the joint distribution of any finite sample from that set. Let {l1, l2, . . . , lnL } = L ⊂ D be of size nL. The nLq ×1 random ector $\boldsymbol{v}_L = (\boldsymbol{v}(\boldsymbol{l}_1)^T, \boldsymbol{v}(\boldsymbol{l}_2)^T, \ldots, \boldsymbol{v}(\boldsymbol{l}_n)^T)^T$ has joint density $p(\boldsymbol{v}_L)$.

IV.CONSTRUCTING SPATIAL MULTIVARIATE DAGS

We now introduce the necessary terminology and notation, which are the basis for later detailing of estimation and prediction algorithms involving SpamTrees. The specifics for building treed DAGs with user-specified depth are in Section 2.1.1, whereas Section 2.1.2 gives details on cherry picking and its use when outcomes are imbalanced and misaligned.

The three key components to build a SpamTree are *(i)* a treed DAG G with *branches* and *leaves* on *M* levels and with depth $\delta \leq M$; *(ii)* a reference set of locations S; *(iii)* a *cherry picking* map. The graph is G = $\{\boldsymbol{V}, \boldsymbol{E}\}$ where the nodes are $\boldsymbol{V} = \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_{mV}\} = \boldsymbol{A} \cup \boldsymbol{B}$, $\boldsymbol{A} \cap \boldsymbol{B} = \emptyset$. We separate the nodes into *reference* $\boldsymbol{A}$ and *non-reference* $\boldsymbol{B}$ nodes, as this will aid in showing that SpamTrees lead to standalone spatial processes in Section 2.2. The *reference* or *branch* nodes are $\boldsymbol{A} = \{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_{mA}\} = \boldsymbol{A}_0 \cup \boldsymbol{A}_1 \cup \cdots \cup \boldsymbol{A}_{M-1}$, where $\boldsymbol{A}_i = \{\boldsymbol{a}_{i,1}, \ldots, \boldsymbol{a}_{i,m_i}\}$ for all $i = 0, \ldots, M-1$ and with $\boldsymbol{A}_i \cap \boldsymbol{A}_j = \emptyset$ if $i \neq j$. The *non- reference* or *leaf* nodes are $\boldsymbol{B} = \{\boldsymbol{b}_1, \ldots, \boldsymbol{b}_{mB}\}$, $\boldsymbol{A} \cap \boldsymbol{B} = \emptyset$.

V.SPAM TREES AS A STANDALONE SPATIAL PROCESS

We define a valid joint density for any finite set of locations in D∗ satisfying the Kolmogorov consistency conditions in order to define a valid process. Enumerate each of the mS reference subsets as Si = {si1 , . . . , sin } where {i1, . . . , ini } ⊂ {1, . . . , nE}, and each of the mU non-reference subsets as Ui = {ci1 , . . . , cin } where {i1, . . . , ini } ⊂ {1, . . . , nC }. Then introduce V = {V1, . . . , VmV } where mV = mS + mU and Vi = Si for i = 1, . . . , mS, VmS +i = Ui fori = 1, . . . , mU . Then take vi = (w(li1), . . . , w(lin))T as the ni × 1 random vector withelements of w(l) for each l ∈ Vi. Denote v[i] = v(η—1(Pa[zi])).

VI.BAYESIAN SPATIAL REGRESSIONS USING SPAMTREES

Suppose we observe an *l*-variate outcome at spatial locations $\boldsymbol{l} \in D \subset \square^d$ which we wish to model using a spatially-varying regression model:

$$y_j(\boldsymbol{l}) = \boldsymbol{x}_j(\boldsymbol{l})^T \boldsymbol{\beta}_j + z_{jk}(\boldsymbol{l}) w(\boldsymbol{l}, \boldsymbol{\xi}_k) + \varepsilon_j(\boldsymbol{l}), \quad j = 1, \ldots, l,$$

where yj(l) is the j-th point-referenced outcome at l, xj(l) is a pj × 1 vector of spatially referenced predictors linked to constant coefficients β , εj(l) i∼id N (0, τ 2) is the measurement.

j

VII.GAUSSIAN SPAM TREES

Enumerate the set of nodes as V = {z1, . . . , zmV }, mV = mS + mU and denote vi = w(η—1(zi)), Cij as the ni × nj covariance matrix between vi and vj, Ci,[i] the ni × Ji covariance matrix

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 07, NCIRTAI - 2026

between vi and v[i], Ci the ni × ni covariance matrix between vi and itself, and C[i] the Ji ×Ji covariance matrix between v[i] and itself.

~

VIII. ESTIMATION AND PREDICTION

I introduce notation to aid in obtaining the full conditional distributions

$$\boldsymbol{y}(\boldsymbol{l}) = \boldsymbol{E}(\boldsymbol{l})\boldsymbol{\beta} + \boldsymbol{E}(\boldsymbol{l})\boldsymbol{v}(\boldsymbol{l}) + \boldsymbol{o}(\boldsymbol{l}),$$

Computing and Storage Cost

The update of τ 2 and β can be performed at a minimal cost as typically p = Σlpj is small; almost all the computation budget must be dedicated to computing p(v | θ) and sampling p(v | y, β, z 2).error for outcome j, and $z_{jk}(\boldsymbol{l})$ is the k-th (of q) covariates for the j-th outcome modeled with spatially-varying coefficient $w(\boldsymbol{l}, \xi_k)$, $\boldsymbol{l} \in \mathrm{D}$, $\xi_k \in \Xi$. For a fixed reference set partition and corresponding nodes, choosing larger δ will result in stronger dependence between leaf nodes and nodes closer to the root—this typically corresponds to leaf nodes being assigned conditioning sets that span larger distances in space. The computational speedup corresponding to choosing $\delta = 1$ can effectively be traded for a coarser partitioning of S, resulting in large conditioning sets that are more local to the leaves.

IX. CONCLUSION AND FUTURE ENHANCEMENT

In conclusion, SPaMTREES provide a scalable and efficient alternative to traditional Gaussian process–based geostatistical models for analyzing large, high-resolution spatial datasets. By leveraging a tree-structured directed acyclic graph and conditional independence assumptions on latent random effects, the approach significantly reduces computational complexity while retaining the ability to capture important spatial and cross-variable dependencies. Guided by information-theoretic principles and efficiency considerations, the model is well-suited for imbalanced and multivariate settings involving data from multiple sources and resolutions. The successful application to both simulated and real-world climate datasets demonstrates its practicality, flexibility, and effectiveness in modeling complex geospatial relationships, making SPaMTREES a promising tool for large-scale spatial data analysis. Future enhancements of SPaMTREES can focus on improving its flexibility, scalability, and applicability to more complex real-world problems. One important direction is extending the model to handle spatio-temporal data, enabling it to capture dynamic changes over time in addition to spatial dependencies. Incorporating online or adaptive learning mechanisms would allow the model to update itself continuously as new data becomes available, making it suitable for real-time monitoring systems. Further improvements in tree construction strategies, potentially using advanced machine learning techniques, could enhance both efficiency and predictive accuracy. Integrating SPaMTREES with deep learning approaches may help in modeling highly nonlinear relationships, especially in high-dimensional datasets such as satellite imagery. Additionally, extending the framework to support non-Gaussian distributions and extreme event modeling would make it more robust for environmental and climate applications. Implementing parallel and distributed computing techniques can further boost scalability for massive datasets. Finally, expanding its application to diverse domains such as urban planning, agriculture, and epidemiology would demonstrate its versatility and broader impact.

X. REFERENCES

[1] S. Ambikasaran, D. Foreman-Mackey, L. Greengard, D. W. Hogg, and M. O'Neil. Fast direct methods for Gaussian processes. IEEE Transactions on Pattern Analysis and Machine Intelligence, 38(2):252–265,2016. doi:10.1109/TPAMI.2015.2448083.

[2] E. Anderson, Z. Bai, C. Bischof, S. Blackford, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum,

[3] S. Banerjee, A. E. Gelfand, A. O. Finley, and H. Sang. Gaussian predictive process models for large spatial data sets. Journal of the Royal Statistical Society, Series B, 70:825–848, 2008. doi:10.1111/j.1467-9868.2008.00663.x.

[4] L. S. Blackford, A. Petitet, R. Pozo, K. Remington, R. C. Whaley, J. Demmel, J. Dongarra, I. Duff,S. Hammarling, G. Henry, et al. An updated set of basic linear algebra subprograms (BLAS).

[5] H. A. Chipman, E. I. George, and R.E. McCulloch. BART: Bayesian additiveregression trees.Annals of Applied Statistics, 4(1):266–298, 2010. doi:10.1214/09-AOAS285.

[6] A. Datta, S. Banerjee, A. O. Finley, N. A. S. Hamm, and M. Schaap. Nonseparable dy- namic nearest neighbor gaussian process models for large spatio-temporal data with an appli-cation to particulate matter analysis. The Annals of Applied Statistics, 10:1286–1316, 2016b. doi:10.1214/16-AOAS931.

[7] M. M. Islam, M. A. Razzaque, M. M. Hassan, W. N. Ismail, and B. Song, "Mobile cloud-based big healthcare data processing in smart cities," IEEE Access, vol. 5, pp. 11 887–11 899, 2017.

[8] S. Bao, M. Chen, and G. Yang, "A method of signal scrambling to secure data storage for healthcare applications," IEEE Journal of Biomedical and Health Informatics, vol. 21, no. 6, pp. 1487–1494, Nov 2017.

[9] M. Akter, A. Gani, M. O. Rahman, M. M. Hassan, A. Almogren, and S. Ahmad, "Performance analysis of personal cloud storage services for mobile multimedia health record management," IEEE Access, vol. 6, pp. 52 625–52 638, 2018.

[10] D. Nychka, S. Bandyopadhyay, D. Hammerling, F. Lindgren, and S. Sain. A multiresolution gaussian process model for the analysis of large spatial datasets. Journal of Computational and Graphical Statistics, 24:579–599, 2015. doi:10.1080/10618600.2014.914946.

[11] Z. C. Quiroz, M. O. Prates, and D. K. Dey. Block Nearest Neighboor Gaussian processes for large datasets, 2016. arXiv:1604.08403

[12] H. Huang, T. Gong, N. Ye, R. Wang, and Y. Dou, "Private and secured medical data transmission and analysis for

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 07, NCIRTAI - 2026

wireless sensing healthcare system," IEEE Transactions on Industrial Informatics, vol.
13, no. 3, pp. 1227–1237, June 2017.

[13] A. Mehmood, I. Natgunanathan, Y. Xiang, H. Poston, and Y. Zhang, "Anonymous authentication scheme for smart cloud based healthcare applications," IEEE Access, vol. 6, pp. 33 552–33 567, 2018.

[14] "Patient home monitoring service leaks private medical data online," https://kromtech.com/blog/security-center/ patient-home-monitoring-service-leaks-private-medical-data-online, 2017.

[15] W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu, "Edge computing: Vision and challenges," IEEE Internet of Things Journal, vol. 3, no. 5, pp. 637–646, Oct 2016.

[16] M. Satyanarayanan, "The emergence of edge computing," Computer, vol. 50, no. 1, pp. 30–39, Jan 2017.

[17] C. Sanderson and R. Curtin. Armadillo: a template-based C++ library for linear algebra. Journal of Open Source Software, 1:26, 2016.

[18] A. V. Vecchia. Estimation and model identification for continuous spatial pro- cesses. Journal of the Royal Statistical Society, Series B, 50:297–312, 2022.

[19] L. Wu, A. Miller, L. Anderson, G. Pleiss, D. Blei, and J. Cunningham. Hierarchical inducing point Gaussian process for inter-domain observations. In Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS), 2023. arXiv:2103.00393.